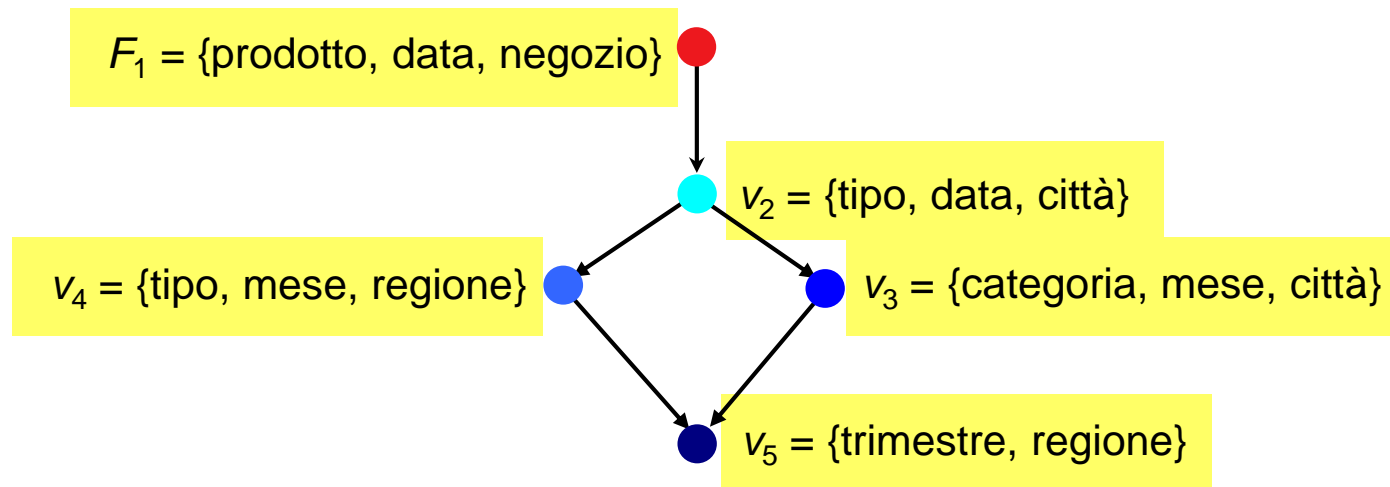


Viste materializzate

Elena Baralis
Politecnico di Torino

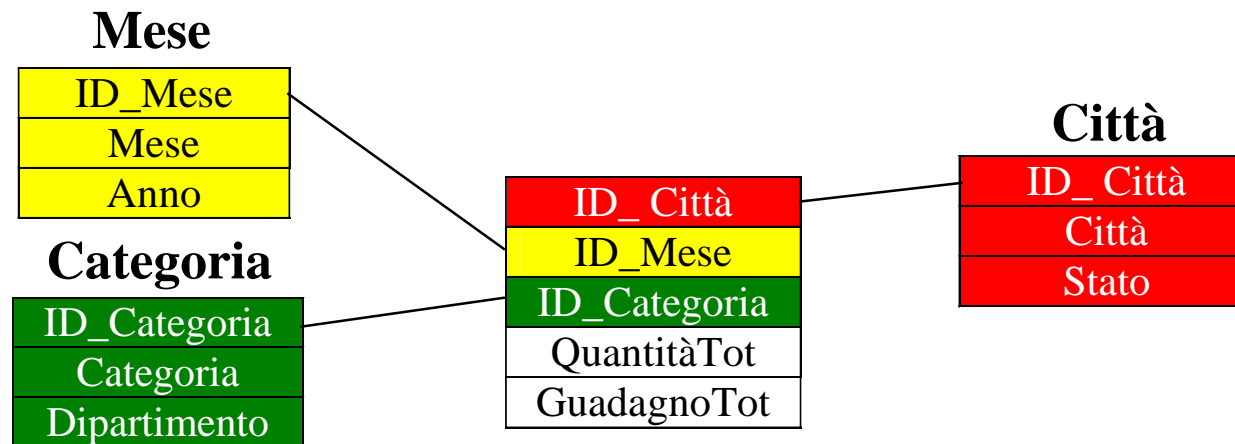
Viste materializzate

- Sommari precalcolati della tabella dei fatti
 - memorizzati esplicitamente nel data warehouse
 - permettono di aumentare l'efficienza delle interrogazioni che richiedono aggregazioni



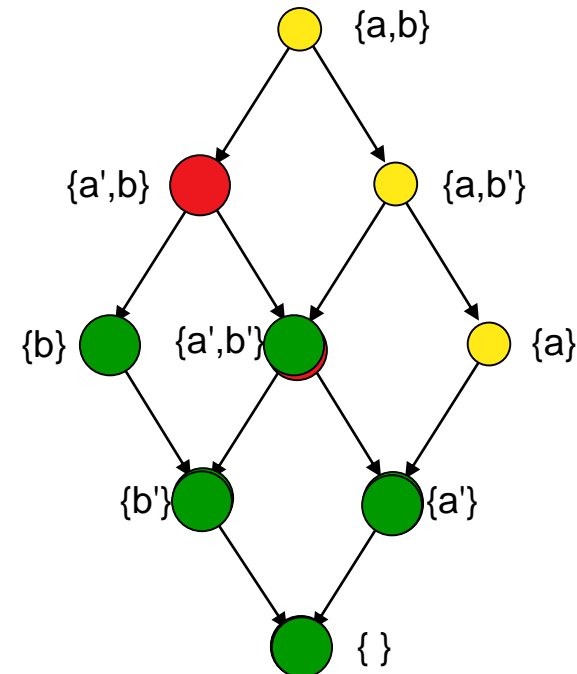
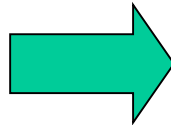
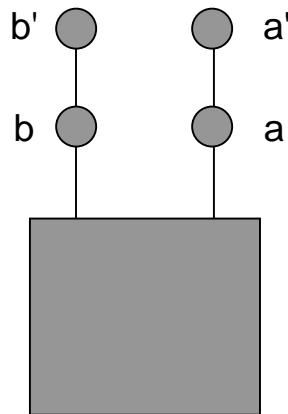
Viste materializzate

- Definite da istruzioni SQL
- Esempio: definizione di v_3
 - a partire da tabelle di base o viste di granularità superiore
 - group by Città, Mese, Categoria
 - aggregazione (SUM) sulle misure Quantità, Guadagno
 - riduzione dettaglio delle dimensioni



Viste materializzate

- Una vista materializzata può essere utilizzata per rispondere a più interrogazioni diverse
 - attenzione al tipo di operatore di aggregazione richiesto

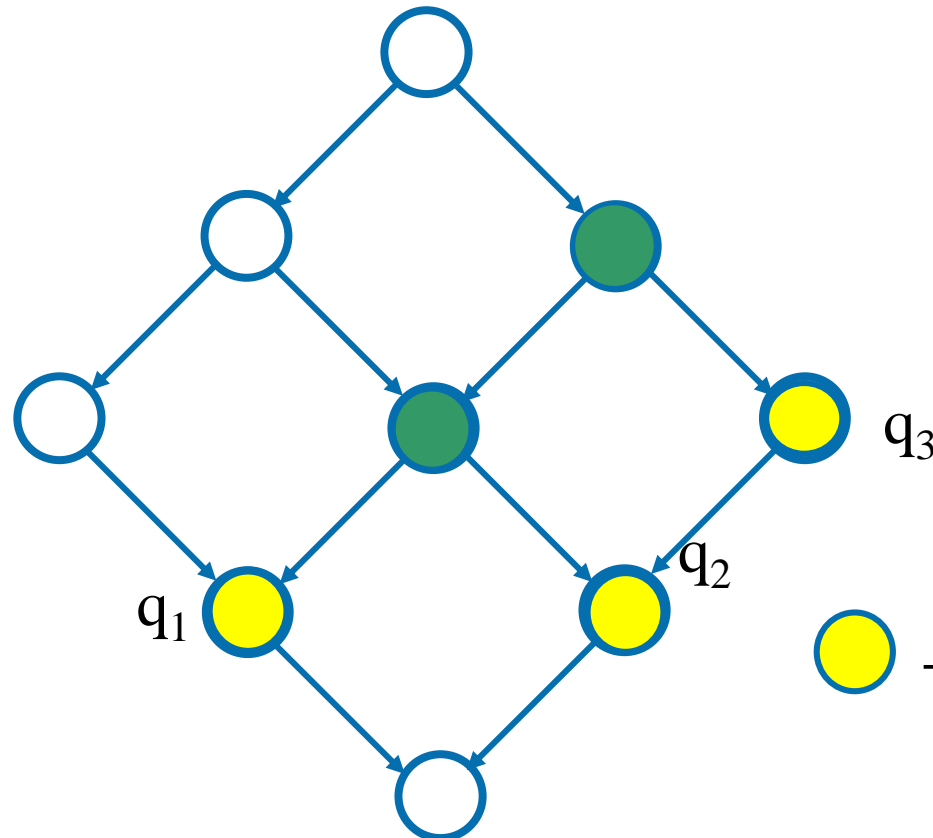




Reticolo multidimensionale

Scelta delle viste

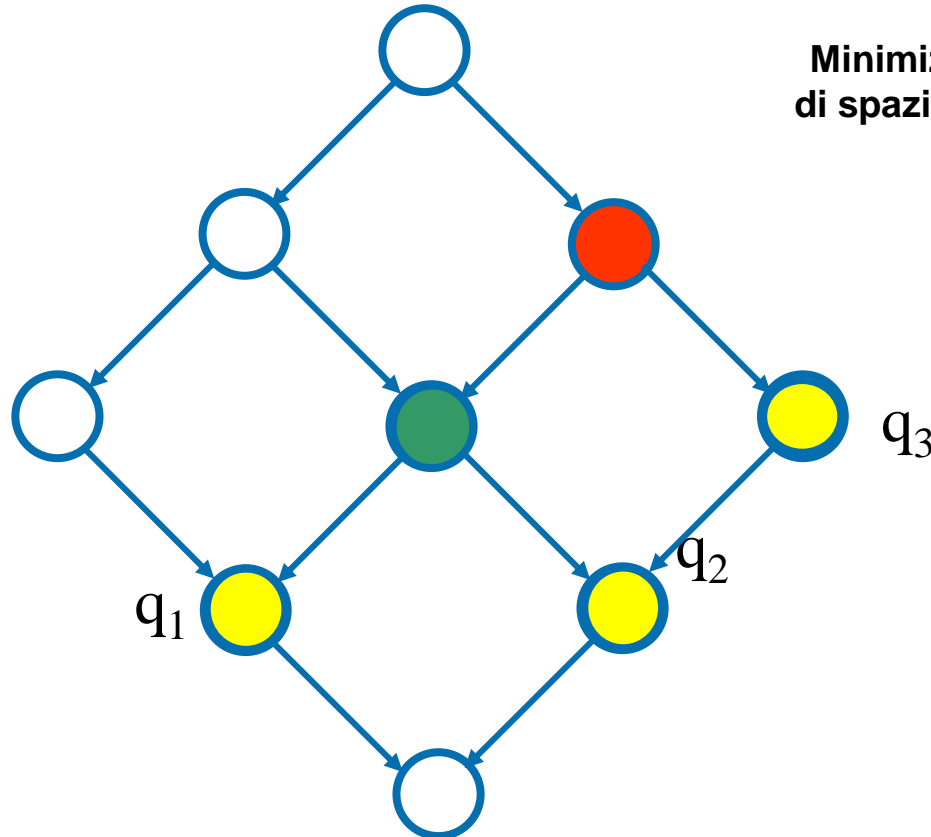
- Numero di possibili combinazioni di aggregazioni molto elevato
 - quasi tutte le combinazioni di attributi sono eleggibili
- Scelta dell'insieme “ottimo” di viste materializzate
- Minimizzazione di funzioni di costo
 - esecuzione delle interrogazioni
 - aggiornamento delle viste materializzate
- Vincoli
 - spazio disponibile
 - tempo a disposizione per l'aggiornamento
 - tempo di risposta
 - freschezza dei dati

Scelta delle viste

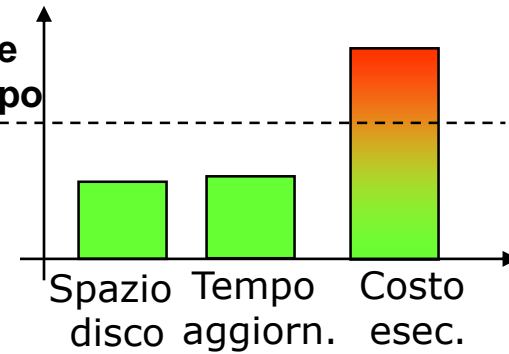


 +  = *viste candidate*,
 ossia potenzialmente
 utili a ridurre il costo
 di esecuzione del
 carico di lavoro

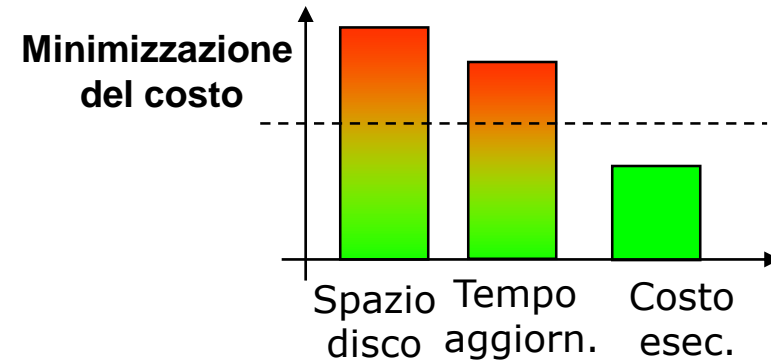
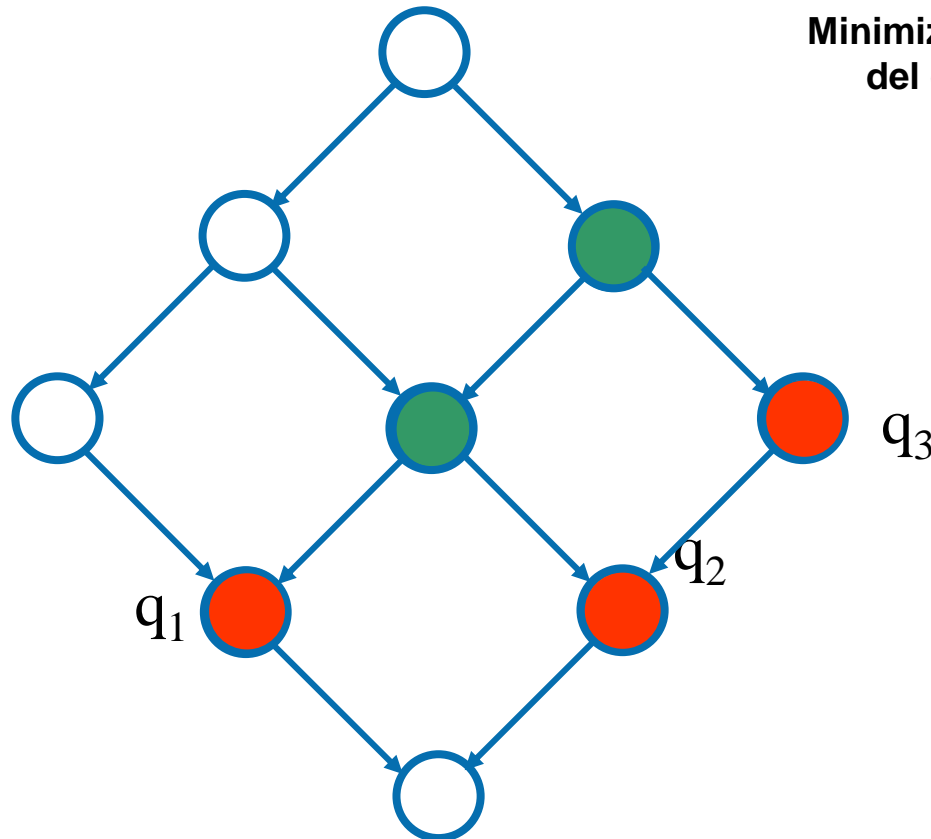
Scelta delle viste



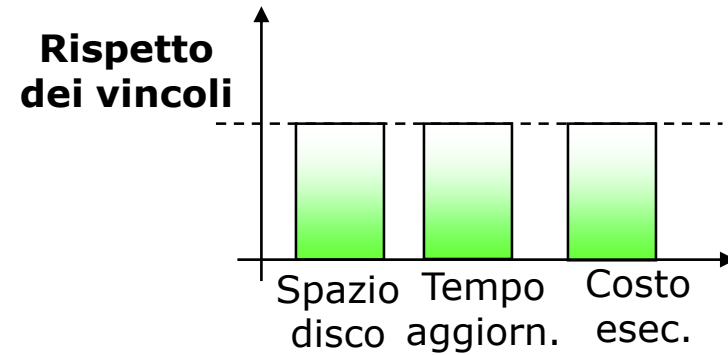
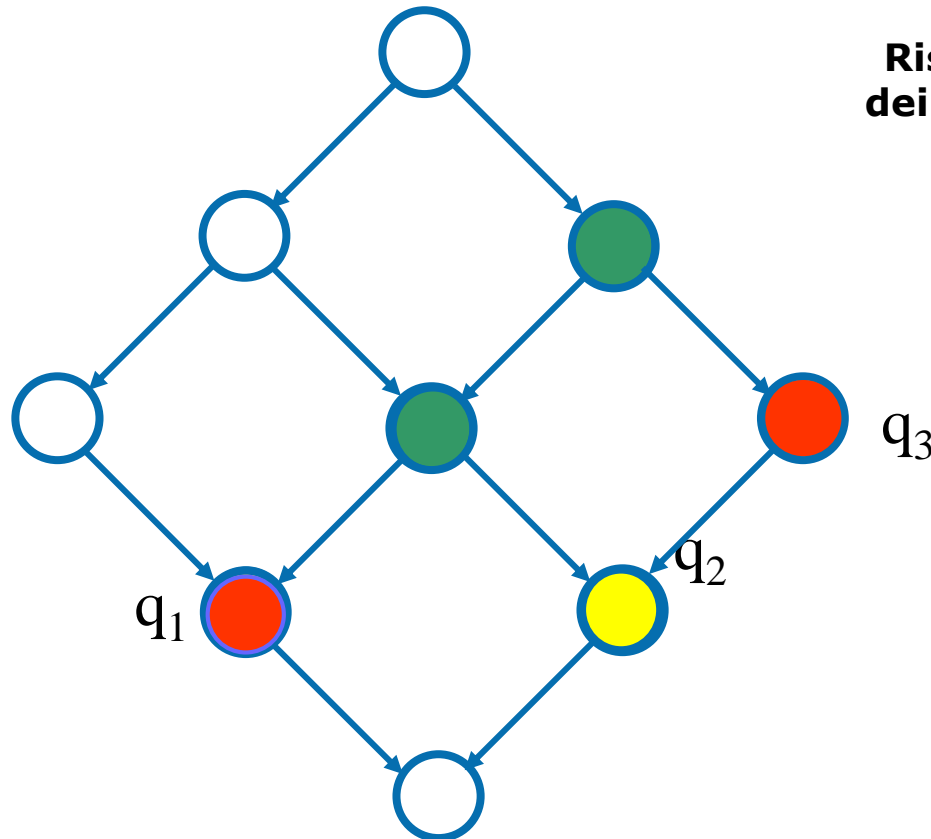
Minimizzazione
di spazio e tempo



Scelta delle viste



Scelta delle viste



Progettazione fisica

Elena Baralis
Politecnico di Torino

Progettazione fisica

- Caratteristiche del carico di lavoro
 - interrogazioni con aggregati che richiedono l'accesso a una frazione significativa di ogni tabella
 - accesso in sola lettura
 - aggiornamento periodico dei dati con eventuale ricostruzione delle strutture fisiche di accesso (indici, viste)
- Strutture fisiche
 - tipologie di indici diverse da quelle tradizionali
 - indici bitmap, indici di join, bitmapped join index, ...
 - l'indice B⁺-tree non è adatto per
 - attributi con dominio a cardinalità bassa
 - interrogazioni poco selettive
 - viste materializzate
 - richiedono la presenza di un ottimizzatore che le sappia sfruttare

Progettazione fisica

- Caratteristiche dell'ottimizzatore
 - deve considerare le statistiche nella definizione del piano di accesso ai dati (cost based)
 - funzionalità di aggregate navigation
- Procedimento di progettazione fisica
 - selezione delle strutture adatte per supportare le interrogazioni più frequenti (o più rilevanti)
 - scelta di strutture in grado di contribuire al miglioramento di più interrogazioni contemporaneamente
 - vincoli
 - spazio su disco
 - tempo disponibile per l'aggiornamento dei dati

Progettazione fisica

- Tuning
 - variazione a posteriori delle strutture fisiche di supporto
 - richiede strumenti di monitoraggio del carico di lavoro
 - spesso necessario per applicazioni OLAP
- Parallelismo
 - frammentazione dei dati
 - parallelizzazione delle interrogazioni
 - inter-query
 - intra-query
 - le operazioni di join e group by si prestano bene all'esecuzione parallela

Scelta degli indici

- Indicizzazione delle dimensioni
 - attributi frequentemente coinvolti in predicati di selezione
 - se il dominio ha cardinalità elevata, indice B-tree
 - se il dominio ha cardinalità ridotta, indice bitmap
- Indici per i join
 - raramente opportuno indicizzare solo le chiavi esterne della tabella dei fatti
 - consigliato bitmapped join index, se disponibile
- Indici per i group by
 - uso di viste materializzate

Alimentazione del data warehouse

Elena Baralis
Politecnico di Torino

Extraction, Transformation and Loading (ETL)

- Processo di preparazione dei dati da introdurre nel data warehouse
 - estrazione dei dati dalle sorgenti
 - pulitura
 - trasformazione
 - caricamento
- semplificato dalla presenza di una staging area
- eseguito durante
 - il primo popolamento del DW
 - l'aggiornamento periodico dei dati

Estrazione

- Acquisizione dei dati dalle sorgenti
- Modalità di estrazione
 - statica: fotografia dei dati operazionali
 - eseguita durante il primo popolamento del DW
 - incrementale: selezione degli aggiornamenti avvenuti dopo l'ultima estrazione
 - utilizzata per l'aggiornamento periodico del DW
 - immediata o ritardata
- Scelta dei dati da estrarre basata sulla loro qualità

Estrazione

- Dipende dalla natura dei dati operazionali
 - storicizzati: tutte le modifiche sono memorizzate per un periodo definito di tempo nel sistema OLTP
 - transazioni bancarie, dati assicurativi
 - operativamente semplice
 - semi-storicizzati: è conservato nel sistema OLTP solo un numero limitato di stati
 - operativamente complessa
 - transitori: il sistema OLTP mantiene solo l'immagine corrente dei dati
 - scorte di magazzino, dati di inventario
 - operativamente complessa

Estrazione incrementale

- Assistita dall'applicazione
 - le modifiche sono catturate da specifiche funzioni applicative
 - richiede la modifica delle applicazioni OLTP (o delle API di accesso alla base di dati)
 - aumenta il carico applicativo
 - necessaria per sistemi legacy
- Uso del log
 - accesso mediante primitive opportune ai dati del log
 - formato proprietario del log
 - efficiente, non interferisce con il carico applicativo

Estrazione incrementale

- Definizione di trigger
 - i trigger catturano le modifiche di interesse
 - non richiede la modifica dei programmi applicativi
 - aumenta il carico applicativo
- Basata su timestamp
 - i record operazionali modificati sono marcati con il timestamp dell'ultima modifica
 - richiede la modifica dello schema della base di dati OLTP (e delle applicazioni)
 - estrazione differita, può perdere stati intermedi se i dati sono transitori

Confronto tra le tecniche di estrazione

	<i>Statica</i>	<i>Marche temporali</i>	<i>Assistita applicazione</i>	<i>Trigger</i>	<i>Log</i>
<i>Gestione dati transitori o semi-storicizzati</i>	NO	Incompleta	Completa	Completa	Completa
<i>Supporto per sistemi basati su file</i>	SI	SI	SI	NO	Raro
<i>Tecnica di realizzazione</i>	Prodotti	Prodotti o sviluppo interno	Sviluppo interno	Prodotti	Prodotti
<i>Costi di sviluppo interno</i>	Nessuno	Medi	Alti	Nessuno	Nessuno
<i>Utilizzo in sistemi legacy</i>	SI	Difficile	Difficile	Difficile	SI
<i>Modifiche ad applicazioni</i>	Nessuna	Probabile	Probabile	Nessuna	Nessuna
<i>Dipendenza delle procedure dal DBMS</i>	Limitata	Limitata	Variabile	Alta	Limitata
<i>Impatto sulle prestazioni del sistema operaz.</i>	Nessuna	Nessuna	Medio	Medio	Nessuna
<i>Complessità delle procedure di estrazione</i>	Bassa	Bassa	Alta	Media	Bassa

Tratto da Devlin, Data warehouse: from architecture to implementation, Addison-Wesley, 1997

Copyright – Tutti i diritti riservati

DATA WAREHOUSE: PROGETTAZIONE - 21

Elena Baralis
Politecnico di Torino

Estrazione incrementale

4/4/2010

Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
3	Barbera	Lumini	75
4	Sangiovese	Cappelli	45

6/4/2010

Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
4	Sangiovese	Cappelli	145
5	Vermentino	Maltoni	25
6	Trebbiano	Maltoni	150

Differenza incrementale

Cod	Prodotto	Cliente	Qtà	Azione
3	Barbera	Lumini	75	D
4	Sangiovese	Cappelli	145	U
5	Vermentino	Maltoni	25	I
6	Trebbiano	Maltoni	150	I

Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

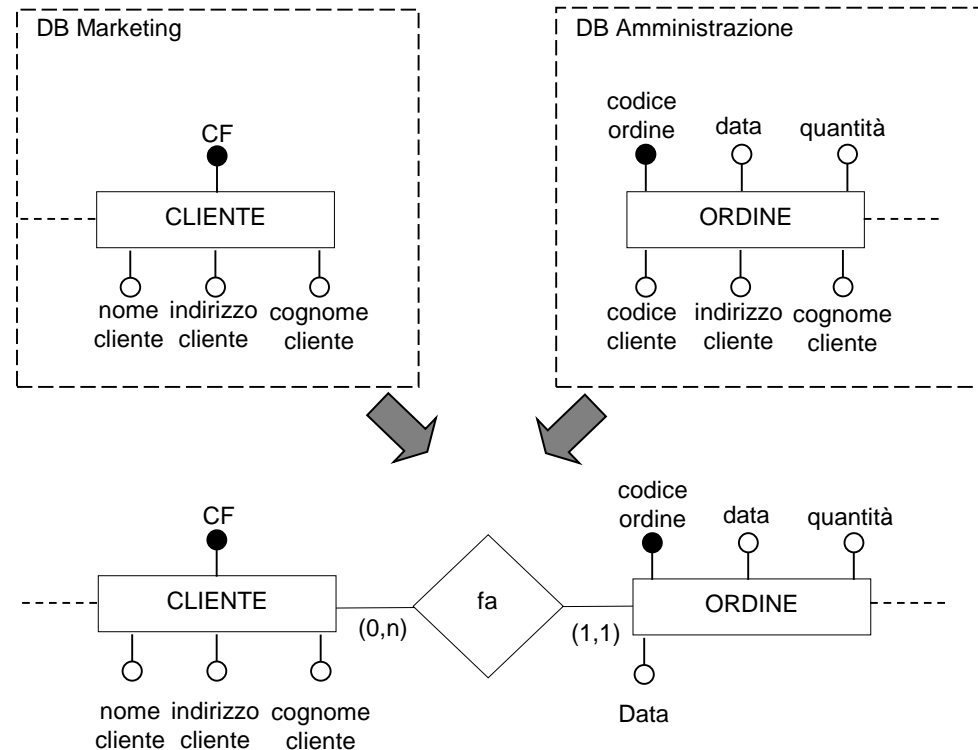
Pulitura

- Operazioni volte al miglioramento della qualità dei dati (correttezza e consistenza)
 - dati duplicati
 - dati mancanti
 - uso non previsto di un campo
 - valori impossibili o errati
 - inconsistenza tra valori logicamente associati
- Problemi dovuti a
 - errori di battitura
 - differenze di formato dei campi
 - evoluzione del modo di operare dell'azienda

Pulitura

- Ogni problema richiede una tecnica specifica di soluzione
 - tecniche basate su dizionari
 - adatte per errori di battitura o formato
 - utilizzabili per attributi con dominio ristretto
 - tecniche di fusione approssimata
 - adatte per riconoscimento di duplicati/correlazioni tra dati simili
 - join approssimato
 - problema purge/merge
 - identificazione di outliers o deviazioni da business rules
- La strategia migliore è la prevenzione, rendendo più affidabili e rigorose le procedure di data entry OLTP

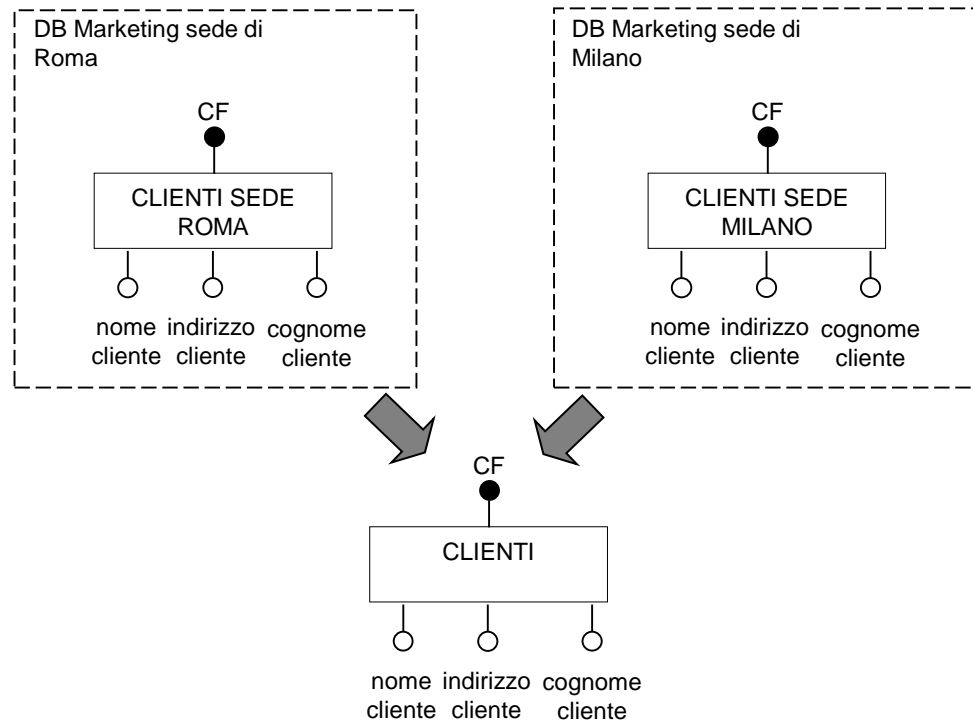
Join approssimato



Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- Il join deve essere eseguito sulla base dei campi comuni, che non rappresentano un identificatore per il cliente

Problema purge/merge



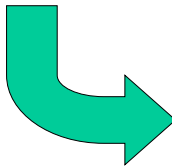
Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- I record duplicati devono essere identificati ed eliminati
- E` necessario un criterio per valutare la somiglianza tra due record

Esempio di pulitura e trasformazione

Elena Baralis
C.so Duca degli Abruzzi 24
20129 Torino (I)

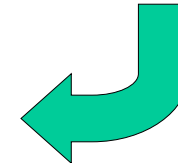
Normalizzazione



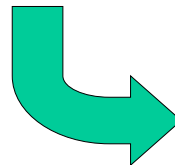
nome:	Elena
cognome:	Baralis
indirizzo:	C.so Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	I

nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	Italia

Standardizzazione



Correzione



nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	10129
città:	Torino
nazione:	Italia

Adattato da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

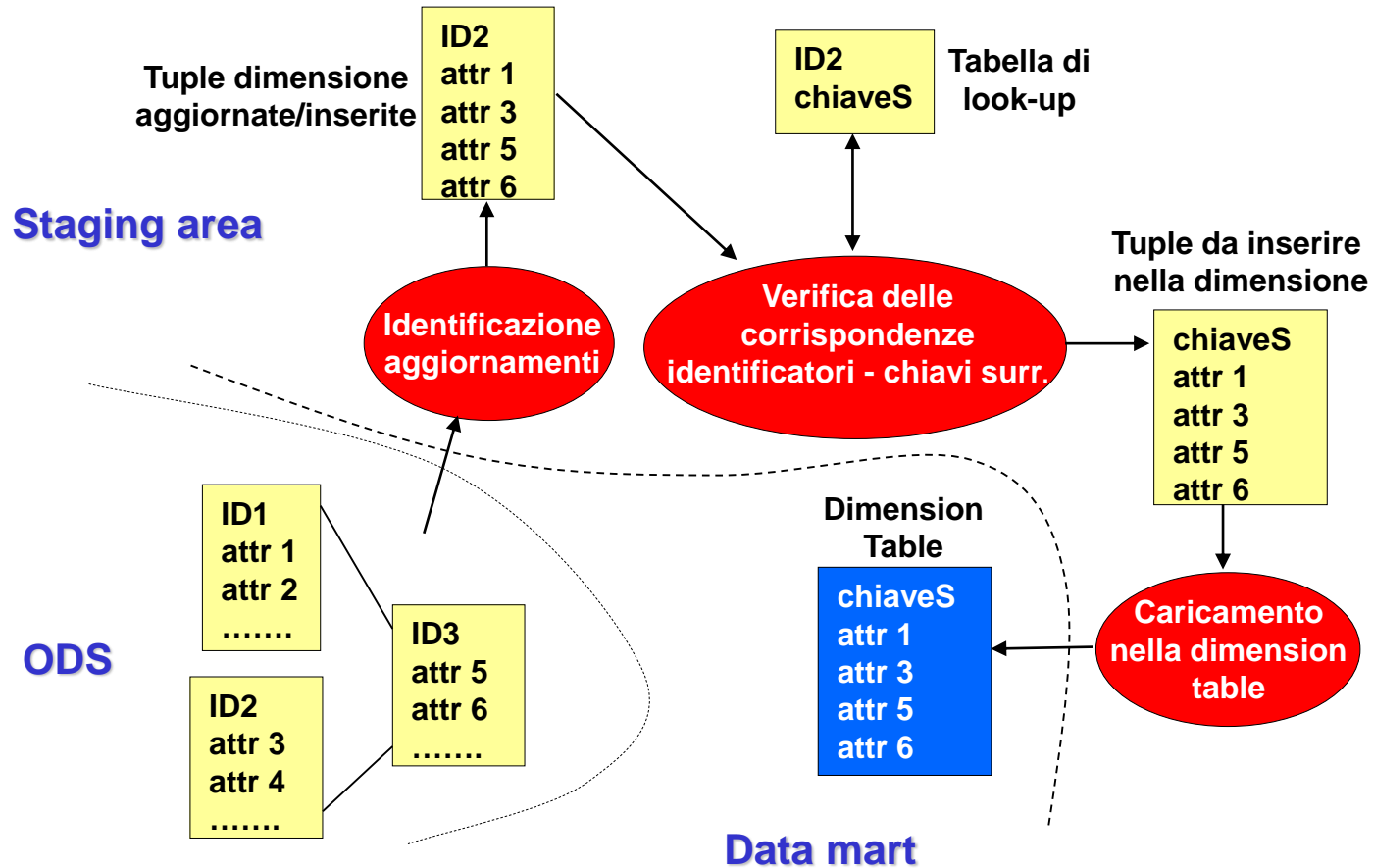
Trasformazione

- Conversione dei dati dal formato operativo a quello del data warehouse (integrazione)
- Richiede una rappresentazione uniforme dei dati operazionali (schema riconciliato)
- Può avvenire in due passi
 - dalle sorgenti operazionali ai dati riconciliati nella staging area
 - conversioni e normalizzazioni
 - matching
 - (eventuale) filtraggio dei dati significativi
 - dai dati riconciliati al data warehouse
 - generazione di chiavi surrogate
 - generazione di valori aggregati

Caricamento

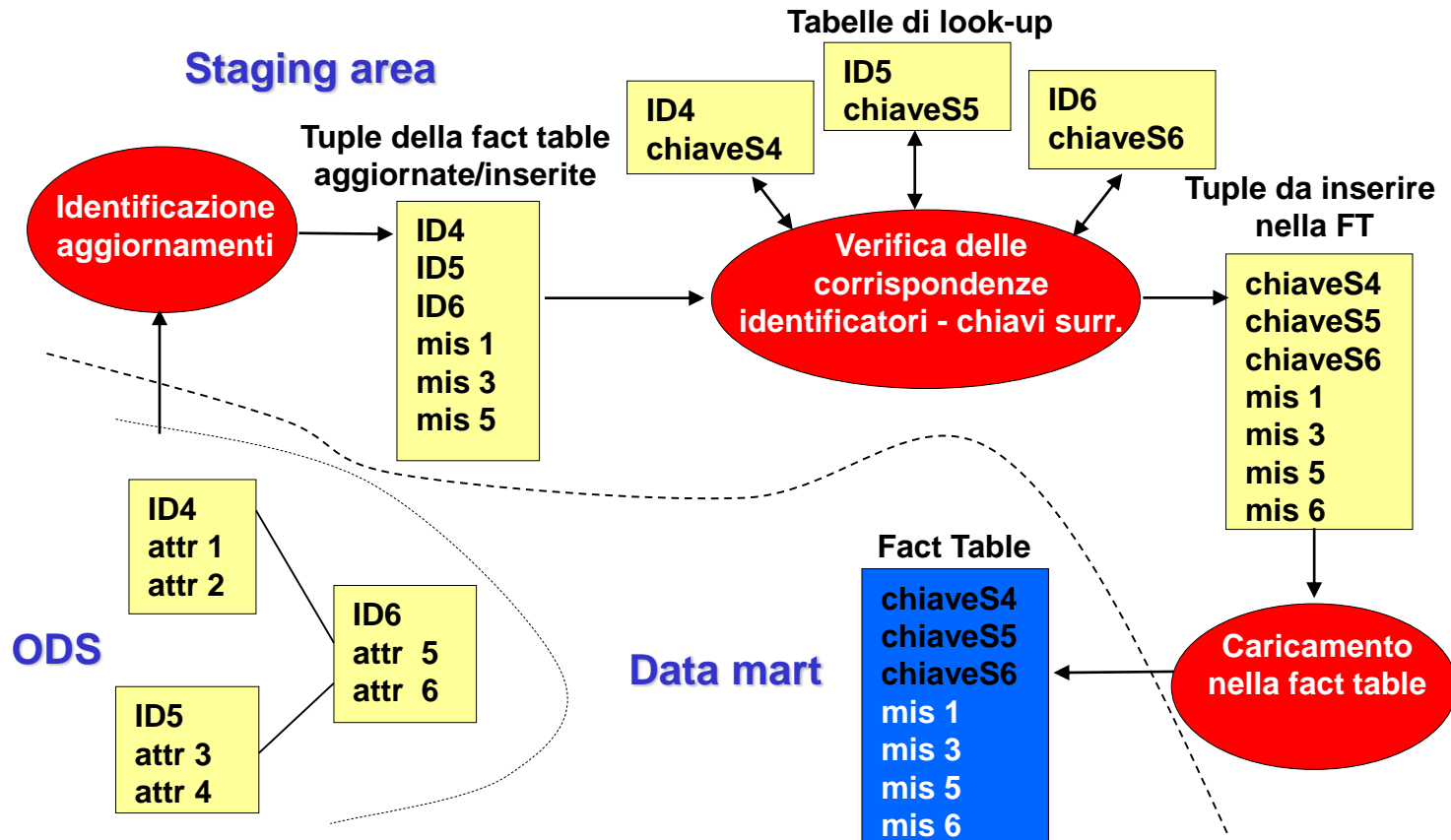
- Propagazione degli aggiornamenti al data warehouse
- Per mantenere l'integrità dei dati, si aggiornano in ordine
 1. dimensioni
 2. tabelle dei fatti
 3. viste materializzate e indici
- Finestra temporale limitata per eseguire gli aggiornamenti
- Richiede proprietà transazionali (affidabilità, atomicità)

Alimentazione delle dimensioni

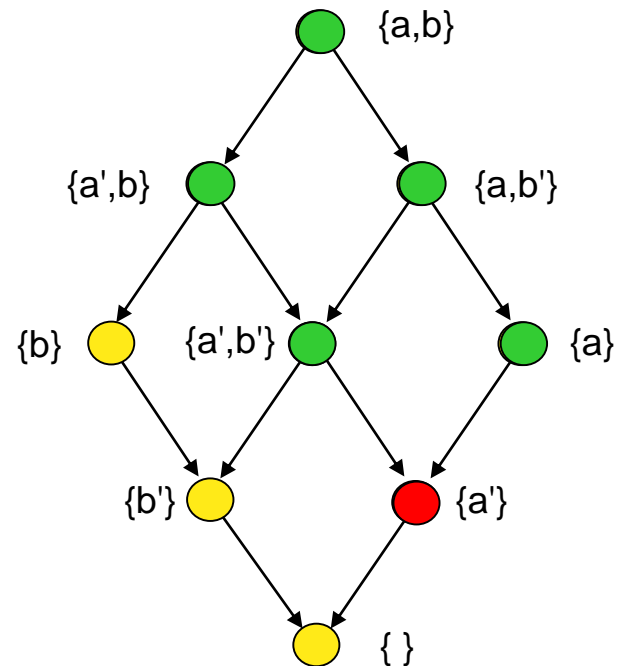


Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Alimentazione delle fact table



Alimentazione delle viste materializzate



Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006